A Beowulf Cluster Machine Equipped with High-Speed Communication Software

Motohiko Tanaka
Institute for Fusion Science, Toki 509-5292, Japan
Email: mtanaka@nifs.ac.jp http://dphysique.nifs.ac.jp/

-- References: NIFS Tech No.12 (May, 2004); Los Alamos Arxiv, physics/ 0407152 (2004).
-- New timing data added: Sep.2005[*1]

Abstract:
A high performance Beowulf (PC cluster) machine installed with Linux operating system and MPI (Message Passing Interface) for interprocessor communication has been constructed using Gigabit Ethernet and a communication software GAMMA (Genoa Active Message Machine) [1] instead of the standard TCP/IP protocol. Fast C/Fortran compilers have been exploited with the GAMMA communication libraries. This method has eliminated large communication overhead of TCP/IP and resulted in significant increase in the computational performance and reasonable scalability on number of processors for real application programs including the first-principle molecular dynamics simulation code*[2].

Keywords: PC cluster machine, non-TCP/IP communication, Gigabit Ethernet, small latency and large throughput, fast C/Fortran compilers

Table 1. Timing of different communication methods using the density-functional ab initio molecular dynamics code Siesta v1.3 [2] for one SCF cycle of a 181 atom system. Four Pentium 4 (3.0GHz), Gigabit Ethernet NIC (3Com996). MPICH is Argonne National Lab's MPI via TCP/IP, and MPI/GAMMA [1] is via non TCP/IP; "FC on" stands for the use of the flow control during data transmission. For comparison, the third row shows the timing with four typical RISC processors - IBM Power 4 (Regatta, 1.5GHz), and the fourth row is the timing for two processors of Dual core Pentium (3.2GHz) under TCP/IP.

		Wallclock a)	CPU time b)	(a)-(b)	(a) / (b)
MPICH TCP/IP		93 sec	67 sec	26 sec	1.39
MPI/GAMMA	FC on FC off	66 sec 115 sec	66 sec 98 sec	0.1sec 17 sec	1.00 1.17
RISC machine (1.5GHz)		59 sec	59 sec	0.1 sec	1.00
EM64/T dual core (3.2GHz) TCP/IP		58 sec	37 sec	21 sec	1.58

References

[1] G.Chiola and G.Ciaccio, Genoa Active Message Machine, http://www.disi.unige.it/project/gamma/

[2] A.Garcia et al., Siesta [Spanish Initiative for Electronic Simulations with Thousands of Atoms], http://www.uam.es/departamentos/ciencias/fismateriac/siesta

Notes:

1. Timing measurement for Dual core Pentium (3.2GHz) x 2 was performed on Sep.1, 2005 under SuSE Linux Enterprise Server 9 (SP-2) and PGI-6.0-4.

2. Construction of a PC cluster machine and the installation of the Siesta code were done under close collaboration with Dr.Y.Zempo (2001-2002). The author thanks Dr.G.Ciaccio for his kind advices on installation of the GAMMA system.